Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery

Stewart, Robert; Auffret, MD; Warr, Amanda; Walker, Alan; Roehe, R; Watson, Mick

Published in:
Nature Biotechnology

DOI:
10.1038/s41587-019-0202-3

Print publication: 02/08/2019

Document Version
Peer reviewed version

Editors summary

Comprehensive resource of cow rumen genomes and a database of predicted proteins.

Compendium of 4941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery

Robert D. Stewart[1], Marc D. Auffret[2], Amanda Warr[1], Alan W. Walker[3], Rainer Roehe[2] and Mick Watson[1*]

[1]The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush EH25 9RG, UK.
[2]Scotland's Rural College, Edinburgh EH25 9RG, UK.
[3]The Rowett Institute, University of Aberdeen, Aberdeen AB25 2ZD, UK
* Corresponding author: mick.watson@roslin.ed.ac.uk

Abstract

Ruminants provide essential nutrition for billions of people worldwide. The rumen is a specialised stomach that is adapted to the breakdown of plant-derived complex polysaccharides. The genomes of the rumen microbiota encode thousands of enzymes adapted to the digestion of plant matter which dominates the ruminant diet. We assembled 4941 rumen microbial metagenome-assembled genomes (MAGs) using ~ 6.5 terabytes of short- and long-read sequence data from 283 ruminant cattle. We present a genome-resolved metagenomics workflow that enabled assembly of bacterial and archaeal genomes that were at least 80% complete. Of note, we obtained 3 single-contig, whole-chromosome assemblies of rumen bacteria, two of which represent previously unknown rumen species, assembled from long-read data. Using our rumen genome collection, we predicted and annotated the largest set of rumen proteins to date. Our set of rumen MAGs increases the rate of mapping of rumen metagenomic sequencing reads from 15% to 50-70%. These genomic and protein resources will enable a better understanding of the structure and functions of the rumen microbiota.

Introduction

Ruminants convert human-inedible, low value plant biomass into products of high nutritional value, such as meat and dairy products. The rumen, which is the first of four chambers of the stomach, contains a mixture of bacteria, archaea, fungi and protozoa that ferment complex carbohydrates e.g. lignocellulose, cellulose, to produce short-chain fatty acids (SCFAs) that the ruminant uses for homeostasis and growth. Rumen microbes are a rich source of enzymes for plant biomass degradation for use in biofuels production[1–3], and manipulation of the rumen microbiome offers opportunities to reduce the cost of food production[4].

Ruminants are important for both food security and climate change. For example, methane is a by-product of ruminant fermentation, released by methanogenic archaea, and an estimated 14% of methane produced by humans has been attributed to ruminant livestock[5]. Methane production has been directly linked to the abundance of methanogenic archaea in the rumen[6], offering possibilities for mitigating this issue through selection[7] or manipulation of the microbiome. Two studies have reported large collections of rumen microbial genomes. Stewart *et al* assembled 913 draft metagenome-assembled genomes (MAGs) (named rumen-uncultured genomes (RUGs)) from the rumen of 43 cattle raised in Scotland[8] and Seshadri *et al* reported 410 reference archaeal and bacterial genomes from the Hungate collection[9]. As isolate genomes, the Hungate genomes are generally higher quality, and crucially, exist in culture, so can be grown and studied in the lab. However, we found that addition of the Hungate genomes only increased read classification by 10%, compared to an increase of 50-70% when the RUGs are used, indicating significant numbers of undiscovered microbes in the rumen.

We present a comprehensive analysis of more than 6.5Tb of sequence data from the rumen of 283 cattle. Our catalog of rumen genomes (named RUG2) includes 4056 genomes that were not present in Stewart *et al*[8], and brings the number of rumen genomes assembled to date to 5845 . We also present a metagenomic assembly of nanopore (MinION) sequencing data (from one rumen sample) that contains at least three whole bacterial chromosomes as single contigs, and which represents the most continuous metagenomic assembly from the rumen to date. These genomic and protein resources will underpin future studies on the structure and function of the rumen microbiome.

Results

4941 metagenome-assembled genomes from the cow rumen

We sequenced DNA extracted from the rumen contents of 283 beef cattle (characteristics of animals sequenced is in Supplementary Data 1), producing over 6.5Tb of Illumina sequence data. We operate a continuous assembly-and-de-replication pipeline, which means that newer genomes of the same strain (>99% average-nucleotide identity, ANI) can replace older genomes if their completeness and contamination statistics are better. All of the 4941 RUGs we present here have completeness >= 80% and contamination <= 10% (Supplementary Figure 1).

4941 RUGs were analysed using MAGpy[10] and their assembly characteristics, putative names and taxonomic classifications are in Supplementary data 2. Sourmash[11], DIAMOND[12] and PhyloPhlAn[13] outputs, which reveal genomic and proteomic similarity to existing public data, are in Supplementary data 3. A phylogenetic tree of the 4941 RUGs, alongside 460 public genomes from the Hungate collection, is presented in Figure 1 and Supplementary data 4. The tree is dominated by large numbers of genomes from the *Firmicutes* and *Bacteroidetes* phyla (dominated by *Clostridiales* and *Bacteroidales* respectively), but also contains many novel genomes from the *Actinobacteria*, *Fibrobacteres* and *Proteobacteria* phyla. *Clostridiales* (2079) and *Bacteroidales* (1081) are the dominant orders, with *Ruminoccocacae* (1111) and *Lachnospiraceae* (640) the dominant families within the *Clostridiales*, and *Prevotellaceae* (521) the dominant family within the *Bacteroidales*.

The genome taxonomy database (GTDB) proposed a new bacterial taxonomy based on conserved concatenated protein sequences[14], and we include the GTDB predicted taxa for all

RUGs (Supplementary data 3). 4763 RUGs have < 99% average-nucleotide-identity (ANI) with existing genomes, and 3535 have < 95% ANI with existing genomes and therefore represent potential novel species.

144 of 4941 genomes are classified to species level, 1092 of 4941 to genus level, 3188 of 4941 to family, 4084 to order, 4514 to class, 4801 to phylum and 4941 to kingdom. Of the genomes classified at the species level, 43 represent genomes derived from uncultured strains of *Ruminococcus flavefaciens*, 42 represent genomes from uncultured strains of *Fibrobacter succinogenes*, 18 represent genomes from uncultured strains of *Sharpea azabuensis*, and 10 represent genomes from uncultured strains of *Selenomonas ruminantium*. These species belong to genera known to play an important role in rumen homeostasis[15].

We assembled 126 archaeal genomes, 111 of which are species of *Methanobrevibacter*. There are two other members of the *Methanobacteriaceae* family, both predicted to be a member of the *Methanosphaera* genus by GTDB. Nine of the archaeal RUGs have sourmash hits to "*Candidatus* Methanomethylophilus sp. *1R26*"; a further three have weak sourmash hits to "Methanogenic archaeon ISO4-H5"; and the remaining archaeal genome has no sourmash hits, and weaker DIAMOND hits to the same genome ("Methanogenic archaeon ISO4-H5"). All thirteen are predicted to be members of the genus "*Candidatus* Methanomethylophilus*"* by GTDB, but this is based on similarity to only two genomes, both of which have uncertain phylogenetic lineages. If "*Candidatus* Methanomethylophilus*"* is a true genus, then our dataset increases the number of sequenced genomes from 2 to 15.

Genome quality statistics were measured by analysing single copy core-genes (Supp Figure 1). There are different standards for the definition of MAG quality: Bowers *et al*[16] describe high-quality drafts as having >= 90% completeness and <= 5% contamination; 2417 of the RUGs meet these criteria. Alternatively, Parks *et al*[17] define a quality score as "Completeness – (5 * contamination)" and exclude any MAG with a score less than 50; 4761 of the RUGs meet that criterion; however, whilst the MAGs from Parks *et al* could be as low as 50% complete, the genomes presented here are all >= 80% complete. The RUGs range in size from 456kb to 6.6Mb with N50s (50% of assembled bases are in contigs greater than the N50 value) ranging from 4.5kb to 1.37Mb. The average number of tRNA genes per RUG is 16.9 and 446 of the RUGs have all 20. As assemblies of Illumina metagenomes struggle to assemble repetitive regions, most of the RUGs do not contain a 16S rRNA gene – 464 RUGs encode a fragment of the 16S rRNA gene, and 154 encode at least one full length 16S rRNA gene.

The coverage of each RUG in each sample is in Supplementary Data 5. Using a cut-off of 1X coverage, most RUGs (4863) are present in more than one animal, 3937 are present in more than 10 animals, and 225 RUGs are present in more than 200 animals. One RUG is present in all animals, RUG11026 a member of the *Prevotellaceae* family.

A near-complete single-contig Proteobacteria genome

Metagenomic assembly of Illumina data often results in highly fragmented assemblies but RUG14498, an uncultured *Proteobacteria* species (genome completeness 87.91% and

contamination 0%) has 136 of 147 single-copy genes present with no duplications in a single contig of just over 1Mb in size. *Proteobacteria* with small genomes (<1.5Mb size) are relatively common (n=67) in our dataset and have also been found in other large metagenome assembly projects[17]. The *Proteobacteria* genomes we present encode proteins with only 45% to 60% amino acid identity with proteins in UniProt TREMBL[18]. We compared our single-contig *Proteobacteria* assembly with nine *Proteobacteria* with similarly sized genomes assembled by Parks *et al*[17] (see whole-genome alignments in Supplementary Figure 2). Average nucleotide identity (ANI; often used to delineate new strains and species) between the 9 UBA genomes and RUG14498 is revealing. UBA2136, UBA1908, UBA3307, UBA3773 and UBA3768 have no detectable level of identity with any other genome in the set; UBA4623, UBA6376, UBA6864, and UBA6830 all share greater than 99.4% average nucleotide identity with one another, indicating that they are highly similar strains of the same species. UBA4623, UBA6376, UBA6864 and UBA6830 also show around 77.8% ANI with RUG14498 suggesting that the single-contig RUG14498 is a high-quality, near-complete whole genome of a novel *Proteobacteria* species. The single contig RUG14498 was assembled by IDBA_ud from sample 10678_020. IDBA_ud exploits uneven depth in metagenomic samples to improve assemblies. RUG14498 is the tenth most abundant genome in 10678_020, and other genomes of similar depth in that sample are taxonomically unrelated, enabling IDBA_ud to assemble almost the entire genome in a single contig.

RUG14498 has a single full length 16S rRNA gene (1507bp). The top hit in GenBank (97% identity across 99% of the length) is accession AB824499.1, a sequence from an "uncultured bacterium" from "the rumen of Thai native cattle and swamp buffaloes". The top hit in SILVA[19] is to the same sequence, only this time annotated as an uncultured *Rhodospirillales*. Together these results support the conclusion that RUG14498 represents a novel *Proteobacteria* species. Low amino acid identity to known proteins limits our ability to predict function and metabolic activity; nevertheless, RUG14498 encodes 73 predicted CAZymes, including 42 glycosyl transferases and 19 glycosyl hydrolases, suggesting a role in carbohydrate synthesis and metabolism.

Novel microbial genomes from the rumen microbiome

We compared 4941 RUGs to the Hungate collection and to our previous dataset[8] (Figure 2). 149/4941 RUGs share > 95% protein identity with Hungate members; 271/4941 > than 90%; this leaves 4670/4941 RUGs with < 90% protein identity with Hungate members. 2387/4941 RUGs have < 90% protein identity with genomes in Stewart *et al*, and more than 1100 RUGs have < 70% protein identity with Stewart *et al*. Many of the RUGs with the lowest protein identity to public genomes could not be classified beyond Phylum level, and some are simply "uncultured bacterium".

We compiled a database comprising all RUG genomes, the Hungate collection genomes[9], and rumen MAGs from Hess *et al*[1], Parks *et al*[17], Solden *et al*[20] and Svartström *et al*[21] that we name the "rumen superset". The rumen superset was dereplicated at both 99% (strain-level) and 95% (species level) average-nucleotide identity (ANI). At 95% ANI, the rumen superset was reduced to 2690 clusters, representing species-level bins. 2078 of these clusters contain only RUG genomes, and therefore represent putative novel rumen microbial species identified in this study. 58 clusters contain both Hungate and RUG genomes, and 268 clusters contained only Hungate genomes (Supplementary Data 6). At 99% ANI, the rumen superset was reduced to 5574 clusters, representing strain-level bins. 4845 of these clusters contain only RUG genomes, and may

represent putative novel rumen microbial strains (Supplementary Data 7). Supplementary Figure 3 shows how the various rumen MAG sets overlap at 95% ANI after de-replication.

We calculated an estimate of the completeness of the RUG2 dataset using the Chao 1 estimator[22] (note we can only do this for our own dataset as it is based on the number of times species are observed at different frequencies, and we do not have these values for other datasets). De-replicating all RUG genomes at 95% gives us 2180 species-level bins. 948 of those are singletons (i.e. observed exactly once), and 410 are doublets (i.e. observed exactly twice). Using the Chao 1 formula, we predict 3276 species, which means we estimate that we have discovered 66.54% of the species present in our samples.

We assessed the impact of using rumen genomic data on the read classification rates of several public datasets using three databases – the first, our custom rumen kraken database consisting of RefSeq complete genomes and the Hungate collection (previously described[23,24]); the second was the same database plus only the RUGs; and the third was the same database plus the rumen superset (which includes the RUGs). We classified five datasets – our own (Stewart *et al*), a dataset we previously published (Wallace *et al*[6]), data from 14 cows from a study on niche specialisation (Rubino *et al*[25]), data from a methane emissions study of sheep (Shi *et al*[26]) and a recent metagenomic study of moose (Svartström *et al*[21]) (Supplementary Figure 4).

The classification rate is increased by using either the RUG or rumen superset databases, though the rumen superset achieves only a marginal increase in most cases. We have improved read classification rates from 15% to 70%, with more than a quarter of our samples achieving a classification rate of 80% or higher. These are comparable with read classification rates for the human microbiome as reported by Pasolli *et al*[27].

Strain-level analysis of methane emissions in sheep

Previously Shi *et al*[26] found no significant changes in community structure between low- and high- methane emitting sheep, although there were differences in gene expression between the two groups. We re-analysed the Shi *et al* dataset using our rumen metagenomic data; specifically, we used our custom kraken database consisting of RefSeq genomes and the rumen superset and used it to classify reads at the level of Kingdom, Phylum, Family, Genus and Species, and tested differences between low methane-emitting (LME) and high-methane emitting (HME) sheep. Whilst we found no significant differences at the level of Kingdom, we found significant and profound differences at every other taxonomic level tested (Supplementary Tables 1-5 and Supplementary Figures 5-9). At the Genus level, *Sharpea*, *Kandleria*, *Fibrobacter* and *Selenomonas* are associated with LME sheep, and *Elusimicrobium* with HME sheep (Supplementary Table 4). At the species level, we found that 340 species differ significantly between LME and HME emitting sheep (Supplementary Table 5), including eleven species of *Bifidobacterium,* and six species of *Olsenella,* all significantly more proportionally abundant in LME sheep, and nine species of *Desulfovibrio* significantly more proportionally abundant in HME sheep. *Fibrobacter succinogenes*, an important rumen microbe known to be heavily involved in plant fibre degradation, is also significantly different between the two groups, and is associated with LME sheep. Some of these microbes were previously identified as differentially proportionally abundant between LME and

HME sheep[15,28] using marker-gene sequencing, though our results provide greater resolution and for the first time reveal the genome sequences involved.

Kraken classifies data at different levels of the NCBI taxonomy; unfortunately, this does not give us data on the RUGs which do not yet have specific NCBI taxonomy IDs. Therefore, to estimate the abundance of individual strains, we aligned reads directly to the rumen superset, and used the number of reads designated as primary alignments as a proxy for the relative abundance of each genome. At FDR<=0.05, 1709 genomes show differentially proportional abundance between low- and high-methane sheep (Supplementary Data 8, Supplementary Figure 10). In supplementary figure 10, LME and HME sheep are clearly separated along principal component 1, which explains 58% of the variance in the data. Supplementary data 8 lists the differentially abundant genomes; of note are large numbers of previously uncharacterised *Lachnospiraceae* species associated with LME sheep; and 22 strains of *Sharpea azabuensis* all higher proportional abundance in LME sheep (all 18 *Sharpea azabuensis* RUGs and four *Sharpea azabuensis* strains from the Hungate collection). These results agree with previous studies based on marker-genes[15], and our dataset increases the number of *Sharpea azabuensis* genomes publicly available from 4 to 22. Large numbers of uncharacterised *Ruminococcaceae* and *Bacteroidia* are also associated with HME sheep. Multiple strains of uncharacterised *Proteobacteria*, including RUG14498 described above, are more proportionally abundant in HME sheep; and *Fibrobacter* strains were almost all associated with LME sheep.

The relationship between proportional abundance of Archaea and methane emissions is not simple. Most archaeal strains are present at similar abundance in LME and HME sheep (Supplementary Data 8). RUGs representing novel strains of *Methanobrevibacter* are often more abundant in HME sheep. The RUG with the most striking proportional abundance is RUG12825, which is likely a member of the *Methanosphaera* genus, and is more abundant in LME sheep. The complex relationship between relative abundance of methanogens and methane emissions may underlie our inability to find significant differences in overall archaeal proportional abundance.

That notwithstanding, these data represent the first strain-level view of methane emissions in sheep to our knowledge, and support and confirm the hypothesis that there are major, fundamental changes in rumen metagenomic relative abundance associated with extremes of low and high methane emissions.

Global rumen census updated

The global rumen census attempted to determine the core rumen microbiome by using 16S rRNA sequencing of rumen samples from 742 individual animals from around the world, comprising eight ruminant species[29]. *Prevotella*, *Butyrivibrio*, and *Ruminococcus*, as well as unclassified *Lachnospiraceae*, *Ruminococcaceae*, *Bacteroidales*, and *Clostridiales* were the dominant rumen bacteria and which may represent a core bacterial rumen microbiome. The same species are abundant in our data (Supplementary Data 5). We also find that many *Proteobacteria* are highly abundant, including *Succinivibrio* (Supplementary Data 5). This is noteworthy because *Proteobacteria* were found to be highly abundant in many of the samples from the rumen census, but were not highlighted as being part of the core rumen microbiome.

To further characterise the proportional abundance of *Proteobacteria* we used the rumen superset database to classify data from this study, Wallace *et al*[6], Rubino *et al*[25], Shi *et al*[15] and

Svartström *et al*[21] (Supplementary Figure 11). *Proteobacteria* are present in all datasets; abundant in cattle datasets, but less so in moose and sheep. Given the high proportional abundance of *Proteobacteria* in many samples, and their consistent presence in all of the samples we tested, we suggest adding *Proteobacteria* to the core bacterial rumen microbiome that was proposed by Henderson *et al*[29].

Long-read assembly of complete bacterial chromosomes

We analysed a single sample (10572_0012) using a MinION sequencer and compare Illumina and MinION assembly statistics in Figure 3. Three flowcells produced 11.4Gb of data with a read N50 of 11,585bp. The mean read length was 6144bp, which is short compared to other reports[30,31]. We attribute this to short DNA fragments and nicks caused by bead-beating step during DNA extraction. We assembled long reads using Canu[32], to form an assembly 178Mb in length with an N50 of 268kb. Regardless of length, Canu predicted 31 of the contigs to be circular. These circular contigs might represent putative plasmids or other circular chromosomes.

One problem with single-molecule sequencing technologies is the presence of post-assembly insertions and deletions (indels)[33]. Canu can correct reads but not enough to remove all indels. Detecting sequencing errors without a ground truth dataset is difficult so we hypothesized that most indels would create premature stop-codons and that gene prediction tools (eg Prodigal[34]) would produce truncated proteins. We examined the ratio between the lengths of predicted proteins and their top-hits in UniProt to estimate indels (Supplementary Figure 12). Although these data indicate multiple errors compared with the Illumina short-read data, we corrected errors by polishing with one round of Nanopolish and two rounds of Racon. We set-up a software pipeline to calculate statistics and produce similar plots for any input genome or metagenome called “IDEEL”.

Statistics for all contigs >= 500kb and all contigs predicted to be circular are in Supplementary data 9. The Nanopore assembly contains several single contigs that we predict are complete, or near-complete, circular whole chromosomes.

Prevotella copri nRUG14950 (tig00000032) is a single contig of 3.8Mb which most closely resembles *Prevotella copri* DSM 18205, and which shows high similarity to RUG14032. *Prevotella copri* nRUG14950 is predicted to be 98.48% complete by CheckM[35], with a contamination score of 2.03%; whereas RUG14032 is estimated to be 96.62% complete and 1.35% contaminated. Comparative alignments between *Prevotella copri* nRUG14950, RUG14032 and *Prevotella copri* DSM 18205 can be seen in Supplementary Figure 13. There is a clear and striking relationship between *Prevotella copri* nRUG14950 and RUG14032. These two genomes, both estimated to be near-complete, were assembled from different samples using different techniques, and sequenced with different sequencing technologies. Our assembly of *Prevotella copri* nRUG14950, with only one contig and estimated to be 98.48% complete, represents the most continuous chromosomal assembly of *Prevotella copri* to date, despite having been assembled from a metagenome.

Selenomonas spp. nRUG14951 is a single contig of 3.1Mb in length, predicted to be circular, and with completeness and contamination statistics of 98.13% and 0.16% respectively. The most similar RUG is RUG10160, sharing a mean of 94% protein identity. RUG10160 is estimated 97.66% complete and 0% contaminated. However, the closest public reference genome is *Selenomonas ruminantium* GACV-9, part of the Hungate collection, which shares only ~64%

protein identity with *Selenomonas* spp. nRUG14951. There exists a good whole-genome alignment between *Selenomonas spp.* nRUG14951 and RUG10160 (Supplementary Figure 14), albeit with some evidence of re-arrangements, and some small sections of the genome that are only captured by the Nanopore assembly.

We also identified *Lachnospiraceae bacterium* nRUG14952, which has a 2.5Mb circular, near-complete genome (95.46%), a second RUG13141 (which has 96% protein identity to nRUG14952) and a more distantly related public reference genome (*Lachnospiraceae bacterium* KHCPX20, 63% protein identity to nRUG14952). The nanopore-assembled genome *Lachnospiraceae bacterium* nRUG14952 contains several genome regions that are absent from RUG13141 (Supplementary Figure 15).

nRUG14951 and nRUG14952 represent entire bacterial chromosomes assembled as single contigs and are the first genome assemblies for these species. The remainder of the nanopore assembly contains highly continuous contigs that represent large portions of previously unsequenced bacterial chromosomes. These results taken together demonstrate the power of long reads for assembling complete, whole chromosomes from complex metagenomes.

To assess the advantage of having complete chromosomal assemblies, we annotated the three nanopore whole genomes and the 3 genomes of their closely related RUGs (Supplementary Data 10). The three complete nanopore genomes contain 5, 7 or 3 full length 16S gene sequences respectively, whereas all three RUGs contain none. In addition, the three nanopore genomes are massively enriched for IS family transposase proteins compared to their RUG counterparts. Transposases are associated with insertion sequences in bacterial genomes, and catalyse the transposition of mobile elements[36]. Finally, in all cases, the nanopore assemblies have more annotated COGs ("clusters of orthologous genes"), suggesting that they have a more complete functional annotation than their short-read counterparts.

A protein database for rumen microbial proteomics

We put together a non-redundant dataset of rumen proteins from the 4941 RUGs and 460 publicly available Hungate collection genomes (10.69 million proteins), following the model of UniRef[37] and clustering the protein set at 100% (9.45 million clusters), 90% (5.69 million clusters) and 50% (2.45 million clusters) identity to form RumiRef100, RumiRef90 and RumiRef50.

To assess the novelty of our dataset at the protein level as compared to other rumen MAG datasets, we took RumiRef100 and added over 900,000 predicted proteins from the rumen superset. We clustered these at 90% identity, which resulted in 6.24 million protein clusters. Of these, 5 million clusters contain at least one RUG protein, 4.74 million contain only RUG proteins, and 3.67 million are singletons containing only RUG proteins.

All 10.69 million predicted proteins from the RUGs have been compared to KEGG[38], 460 public genomes from the Hungate collection, UniRef100, UniRef90 and UniRef50. The mean protein identity of the top hit for these databases is 55.88%, 63.58%, 67.52%, 67.25% and 59.97% respectively. These data provide the most comprehensive and richly annotated protein dataset from the rumen to date.

The RUG proteins were compared to the CAZy[39] database (31st July 2018) using dbCAN2[40]. 442,917 are predicted to be involved in carbohydrate metabolism, including 235,001 glycoside

hydrolases, 120,494 glycosyltransferases, 55,523 carbohydrate esterases, 23,928 proteins with carbohydrate binding modules, 6834 polysaccharide lyases, 907 proteins with predicted auxiliary activities, 80 proteins with a predicted cohesin domain, and 150 proteins with an S-layer homology module (SLH).

The similarity of the predicted CAZymes to the current CAZy database can be seen in Figure 4. None of the eight classes of carbohydrate active enzymes displays an average protein identity greater than 60% indicating that CAZy poorly represents the diversity of CAZymes encoded in the genomes of ruminant microbes. Of particular note is the class AA "auxiliary activities", with an average protein identity of less than 30% between CAZy and the RUG CAzymes. AA was created by CAZy to classify ligninolytic enzymes and lytic polysaccharide mono-oxygenases (LPMOs).

The distribution of CAZymes across 12 different phyla and the group of "unknown" bacteria can be seen in Figure 5. The *Bacteroidetes* (3.9 million) and *Firmicutes* (5.3 million) together contribute the largest number of proteins to our dataset; however, whereas 5.7% of the proteome of *Bacteroidetes* is devoted to CAZyme activity, in *Firmicutes* the figure is 3.2%. *Fibrobacteres* devote the highest percentage of their proteome to carbohydrate metabolism (over 6.6%), as is expected due to their fibre-attached, high cellulolytic activity. Only a few studies exist on the role of *Planctomycetes* in the rumen[24,41,42], however whilst they contribute a relatively low number of proteins in our dataset (30172), just over 5% of those proteins are predicted to be CAZymes, suggesting a role in and adaptation to carbohydrate metabolism. 79 out of 80 cohesin-containing proteins are encoded by the *Firmicutes* (the remaining one is encoded by an unknown bacterium), as are 101 out of 149 SLH-domain containing proteins. Both are components of cellulosomes, multi-enzyme complexes involved in fibre degradation, which are encoded by some members of the *Clostridiales* family.

There are 1707 *Bacteroidetes* genomes in the RUGs, and additionally we have a whole genome of *Prevotella copri* from the Nanopore assembly. These 1708 genomes were subject to prediction of polysaccharide utilisation loci (PUL) using our pipeline PULpy[43]. Of the 1708 genomes, 1469 are predicted to have at least one PUL, and in total there are 15,629 separate loci involving 88260 proteins. The highest number of PUL per genome are 52 PUL for RUG13980 and 50 for RUG10279, both labelled uncultured *Prevotellaceae;* both of these genomes are closely related to *Prevotella multisaccharivorax*, known to be able to utilise multiple carbohydrate substrates[44].

Discussion

The rumen microbiome has a crucial role in food security and climate change. Recent studies have released more than 1300 draft and complete rumen genomes. We add 4941 near-complete, de-replicated metagenome-assembled genomes to these 1300 existing rumen genomes[9,20,21]. By combining our dataset with publicly available genomes, we assembled a "rumen superset" of 5845 public bacterial and archaeal genomes. This set contains 2690 unique species-level bins (95% ANI) and 2078 of these 2690 putative species are RUG2 genomes discovered in this study. The RUG2 dataset and the rumen superset bring read classification rates up to 70% for our own data, and 45-55% for other rumen metagenome datasets (some from non-cattle ruminants). The remaining reads are likely to derive from low-abundance bacterial and archaeal species, difficult-to-assemble genomes, and the fungal, protozoan and viral genomes that are not part of this study.

We estimate that we have discovered 65% of rumen species in our samples, representing 4 important beef cattle breeds, which suggests that there are over 1000 species yet to be sequenced and assembled. Given that average read classification rates dip from 70% in our own data, to 50% in the Rubino *et al* cattle data (Limousin x Friesian cross)[25] and Shi *et al* sheep data[26], and 45% in moose[21], there are many species yet to be discovered, and there are likely to be species- and breed- specific rumen microbiomes. We note the high abundance of unclassified *Proteobacteria* in our data, and in the rumen census data, and suggest that these may form part of a core rumen microbiome. Our dataset contains 74 proteobacterial genomes, and we present one near-complete genome in a single contig.

We apply our dataset to re-analyse data on methane emissions in sheep that was published in 2014[26]. Using a combined database of rumen microbial genomes we reveal fundamental and large-scale changes in rumen metagenomic abundance between LME and HME sheep. These differences occur at almost every taxonomic level tested, and the rumen superset database enables us to analyse these data at previously unprecedented resolution. Whilst species- and strain- level metagenomic data must always be interpreted with care – there remains a possibility that strains that are not present in the database are driving the observed differences – nonetheless we observe consistent patterns suggesting large changes in abundance for numerous species. Our analysis confirms and strengthens subsequent studies of methane emissions in sheep[15,28] by identifying specific strains of bacteria and archaea involved and revealing their genome sequence. Our analysis confirms that there is a complex relationship between archaeal abundance and methane emissions, with archaeal species and strains both positively and negatively associated with methane emissions. These insights into metagenomic species- and strain- level aspects of methane emissions will form the basis of future studies.

The main rumen functions rely on the activity of proteins encoded in rumen microbe genomes, and as researchers produce more proteomic data, it is vital that protein reference datasets are available. We present the largest redundant and non-redundant rumen microbial protein prediction datasets to date, and provide rich annotation using public protein, pathway and enzyme databases. This resource will enable researchers to predict the function of each protein, and better assess the functional consequences of changes in the rumen proteome.

Going forwards, it is vital that more rumen bacteria and archaea are brought into culture, to better enable studying the functions of the rumen microbiome. In particular, if we are to design rational interventions to manipulate rumen feed-conversion or methane emissions, we will need to understand microbiome structure, which substrates are utilized by microbiota, and how they interact with one another and the ruminant host. Sequencing and assembling rumen microbial genomes is an important step towards improved culture collections and future manipulation of the rumen microbiome for human benefit.

Competing interests.

The authors declare no competing interests.

Acknowledgements

The Rowett Institute and SRUC are core funded by the Rural and Environment Science and Analytical Services Division (RESAS) of the Scottish Government. The Roslin Institute forms part of

the Royal (Dick) School of Veterinary Studies, University of Edinburgh. This project was supported by the Biotechnology and Biological Sciences Research Council (BBSRC; BB/N016742/1, BB/N01720X/1), including institute strategic programme and national capability awards to The Roslin Institute (BBSRC: BB/P013759/1, BB/P013732/1, BB/J004235/1, BB/J004243/1); and by the Scottish Government as part of the 2016–2021 commission.

Data Availability

Raw sequence reads for all samples are available under ENA project PRJEB31266, except for 10572 which are available under PRJEB21624. All metagenomic assemblies and RUGs are in the process of being deposited in ENA under accession PRJEB31266. All protein predictions, clusters and annotation are available at DOI: 10.7488/ds/2470.

Code Availability

Comparative genomic analysis was carried out using MAGpy[10] (https://github.com/WatsonLab/MAGpy); analysis of PUL was carried out using PULpy[43] (https://github.com/WatsonLab/PULpy); analysis of indels in nanopore data was carried out using IDEEL (https://github.com/mw55309/ideel)

Figure 1 Phylogenetic tree of 4941 rumen uncultured genomes (RUGs) from the cow rumen, additionally incorporating rumen genomes from the Hungate collection. The tree was produced from concatenated protein sequences using PhyloPhlAn[13] and subsequently drawn using GraPhlAn[45]. Labels show Hungate genome names, chosen to be informative but not overlap.

Figure 2 A comparison of the 4941 RUG dataset with the Hungate collection (A) and our previously published data from Stewart et al (B). Black line is average percentage protein identity with closest match (right-hand y-axis), and blue dots are mash distance (k=100,000) between RUG and the closest match (a measure of dissimilarity between two DNA sequences). As expected, a high protein identity relates to a low mash distance, and vice versa. The RUGs are sorted independently for figures A and B, by average protein identity. There is a clear inflection point in Figure 5B, roughly half way along the x-axis, where the protein identity dips below 90% and the Mash distance rises, neatly demonstrating the novelty represented by our new, larger dataset

Figure 3 A comparison of Illumina and Nanopore metagenomic assembly statistics. The coloured histograms show the distribution of statistics for 282 Illumina assemblies, and the single Nanopore assembly is highlighted. A) N50; B) total length of the assembly; and C) length of the longest contig. As can be seen, the Nanopore assembly N50 of 268kb is over 56-times longer than the average Illumina assembly (4.7kb); whilst the Illumina assemblies are often longer (average 600Mb), the Nanopore assembly (at 178Mb in length) is not the shortest of the assemblies we produced; and the Nanopore assembly produced the longest contig at 3.8Mb, seven-times longer than the average for the Illumina assemblies (479kb) and 2.74-times longer than the longest single Illumina contig (1.38Mb – one of thirteen contigs from the 99.19% complete “uncultured *Bacteroidia* bacterium RUG14538”). In terms of a direct comparison, the Illumina-only assembly of the same sample has an N50 of 12.2Kb, a total length of 247Mb and a longest contig of 358Kb

Figure 4 Maximum percentage identity between CAZyme-predicted proteins from the RUGs and the CAZy database. GH=glycoside hydrolase (n=235,001); GT=glycosyl transferase (n=120,494); PL=polysaccharide lyase (n=6,834); CE=carbohydrate esterase (n=55,523); AA=auxiliary activities; CBM=carbohydrate binding module (n=23,928); SLH=S-layer homology domain (n=150); cohesin=cohesin domain (n=80). Centre line shows the median value; box shows the interquartile range; whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box

Figure 5 Top: Total number of proteins for 12 phyla and the group of unknown bacteria; Middle: percentage of the proteome predicted to be CAZymes; Bottom: distribution of eight CAZyme classes as a proportion of the total number of predicted CAZymes. GH=glycoside hydrolase; GT=glycosyl transferase; PL=polysaccharide lyase; CE=carbohydrate esterase; AA=auxiliary activities; CB=carbohydrate binding module; SL=S-layer homology domain; co=cohesin domain

References

1. Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science (80-.).* **331**, 463–467 (2011).

2. Cowan, D. A. *et al.* Metagenomics, gene discovery and the ideal biocatalyst. *Biochem. Soc. Trans.* **32**, 298–302 (2004).

3. Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I. & Watson, M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Front. Genet.* **8**, 23 (2017).

4. Huws, S. A. *et al.* Addressing Global Ruminant Agricultural Challenges Through Understanding the Rumen Microbiome: Past, Present, and Future. *Front. Microbiol.* **9**, 2161 (2018).

5. Gerber, P. J. & Food and Agriculture Organization of the United Nations. *Tackling climate change through livestock : a global assessment of emissions and mitigation opportunities*. (Rome: Food and Agriculture Organization of the United Nations (FAO)., 2013).

6. Wallace, R. J. *et al.* The rumen microbial metagenome associated with high methane production in cattle. *BMC Genomics* **16**, 839 (2015).

7. Roehe, R. *et al.* Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance. *PLoS Genet.* **12**, e1005846 (2016).

8. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 870 (2018).

9. Seshadri, R. *et al.* Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36**, 359–367 (2018).

10. Stewart, R. D., Auffret, M., Snelling, T. J., Roehe, R. & Watson, M. MAGpy: a reproducible pipeline for the downstream analysis of metagenome-assembled genomes (MAGs). *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty905

11. Brown, C. T. & Irber, L. sourmash: a library for MinHash sketching of DNA. *J. Open Source Softw.* **1**, 27 (2016).

12. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

13. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 2304 (2013).

14. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996 (2018).

15. Kamke, J. *et al.* Rumen metagenome and metatranscriptome analyses of low methane yield sheep reveals a Sharpea-enriched microbiome characterised by lactic acid formation and utilisation. *Microbiome* **4**, 56 (2016).

16. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

17. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).

18. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

19. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-6 (2013).

20. Solden, L. M. *et al.* Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nat. Microbiol.* **3**, 1274–1284 (2018).

21. Svartström, O. *et al.* Ninety-nine de novo assembled genomes from the moose (Alces alces) rumen microbiome provide new insights into microbial plant biomass degradation. *ISME J.* **11**, 2538–2551 (2017).

22. Chao, A. Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics* **11**, 265–270 (1984).

23. Auffret, M. D. *et al.* Identification, Comparison, and Validation of Robust Rumen Microbial Biomarkers for Methane Emissions Using Diverse Bos Taurus Breeds and Basal Diets. *Front. Microbiol.* **8**, 2642 (2018).

24. Auffret, M. D. *et al.* The rumen microbiome as a reservoir of antimicrobial resistance and pathogenicity genes is directly affected by diet in beef cattle. *Microbiome* **5**, 159 (2017).

25. Rubino, F. *et al.* Divergent functional isoforms drive niche specialisation for nutrient acquisition and use in rumen microbiome. *ISME J.* **11**, 932–944 (2017).

26. Shi, W. *et al.* Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. *Genome Res.* **24**, 1517–1525 (2014).

27. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20 (2019).

28. Kittelmann, S. *et al.* Two Different Bacterial Community Types Are Linked with the Low-Methane Emission Trait in Sheep. *PLoS One* **9**, e103171 (2014).

29. Henderson, G. *et al.* Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci. Rep.* **5**, 14567 (2015).

30. Risse, J. *et al.* A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. *Gigascience* **4**, 60 (2015).

31. Ip, C. L. C. *et al.* MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research* **4**, 1075 (2015).

32. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

33. Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* **37**, 124–126 (2019).

34. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

35. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–55 (2015).

36. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891 (2014).

37. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–32 (2015).

38. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

39. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233-8 (2009).

40. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).

41. Hook, S. E. *et al.* Impact of subacute ruminal acidosis (SARA) adaptation and recovery on the density and diversity of bacteria in the rumen of dairy cows. *FEMS Microbiol. Ecol.* **78**, 275–284 (2011).

42. Kasparovska, J. *et al.* Effects of Isoflavone-Enriched Feed on the Rumen Microbiota in Dairy Cows. *PLoS One* **11**, e0154642 (2016).

43. Stewart, R. D., Auffret, M., Roehe, R. & Watson, M. Open prediction of polysaccharide utilisation loci (PUL) in 5414 public Bacteroidetes genomes using PULpy. *bioRxiv* 421024 (2018). doi:10.1101/421024

44. Sakamoto, M., Umeda, M., Ishikawa, I. & Benno, Y. Prevotella multisaccharivorax sp. nov., isolated from human subgingival plaque. *Int. J. Syst. Evol. Microbiol.* **55**, 1839–1843 (2005).

45. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015).

Online Methods

*Metagenomic samples***.** Animal experiments were conducted at the Beef and Sheep Research Centre of Scotland's Rural College (SRUC). The experiment was approved by the Animal Experiment Committee of SRUC and was conducted in accordance with the requirements of the UK Animals (Scientific Procedures) Act 1986.

The data were obtained from three cross breeds: Aberdeen Angus, Limousin and Charolais and one pure breed: Luing (Supplementary Data 4). As previously described, the animals were slaughtered in a commercial abattoir where two post-mortem digesta samples (approximately 50 mL) were taken immediately after the rumen was opened to be drained[46,47]. DNA extraction was carried out following the protocol of Yu and Morrison[48] and based on repeated bead beating plus column filtration. Illumina TruSeq libraries were prepared from genomic DNA and sequenced on an Illumina HiSeq 4000 by Edinburgh Genomics.

We experienced severe problems when using the MinION for rumen microbiome DNA when following standard, recommended protocols, and we hope our adapted methods will be of

assistance to others. We found that the DNA did not meet the recommended purity for Nanopore library prep following extraction, according to Nanodrop O/D ratios. RNase treatment using Riboshredder and clean up with methods such as AMPure XP beads were sufficient to obtain O/D ratios within the recommended range, but DNA from these methods typically led to poor or failed sequencing runs. Successful clean-up reaching recommended O/D ratios and leading to successful sequencing runs was carried out using RNase treatment with Riboshredder and a phenol-chloroform purification. 1D libraries were prepared starting with 2ug of DNA per library following Oxford Nanopore's SQK-LSK108 1D ligation protocol with modifications. The incubation in the end prep stage of the protocol was extended to 30 minutes at 20°C and 30 minutes at 65°C and the incubation in the ligation stage was extended to 15 minutes at room temperature. The optional FFPE repair step was also carried out. Three sequencing runs were carried out using FLOMIN-106 flow cells on a MinION MK1b housed in the Watson lab, U. Edinburgh.

Bioinformatics – metagenomic assembly and binning. In total, 282 samples were sequenced for this study generating between 24 and 140 million 150bp paired-end reads per sample. The samples were sequenced in five batches of 48 samples and one batch of 42 samples (this 42-sample batch was the sole basis of Stewart *et al*). An additional sample was used for Hi-C sequencing in Stewart *et al*[8], and the metagenome-assembled genomes from that sample are included in the de-replicated set.

Unless otherwise stated, all parameters used were the default. Each sample was assembled and binned individually using coverage and content as previously described[8]. Briefly, each sample was assembled using idba_ud[49] (v1.1.3) with the options --num_threads 16 --pre_correction --min_contig 300. BWA MEM[50] (v0.7.15) was used to map reads back to the filtered assembly and Samtools[51] (v 1.3.1) was used to convert to BAM format. Script jgi_summarize_bam_contig_depths from the MetaBAT2[52] (v 2.11.1) package was used to calculate coverage from the resulting BAM files. A co-assembly was also produced for each of the 6 batches of samples using MEGAHIT[53] (v1.1.1) with options --kmin-1pass, -m 60e +10, --k-list 27,37,47,57,67,77,87, --min-contig-len 1000, -t 16.

Metagenomic binning was applied to both single-sample assemblies and the coassemblies using MetaBAT2[52], with options --minContigLength 2000, --minContigDepth 2. Single-sample binning produced a total of 37153 bins, and co-assembly binning produced a further 23335. All 60743 bins were aggregated and then dereplicated using dRep[54] (v1.1.2). The dRep dereplication workflow was used with options dereplicate_wf -p 16 -comp 80 -con 10 -str 100 –strW 0. Thus, in pre-filtering, only bins assessed by CheckM (v1.0.5) as having both completeness ≥ 80%, and contamination ≤ 10% were retained for pairwise dereplication comparison (n=10586). Bin scores were given as completeness - 5*contamination + 0.5*log(N50), and only the highest scoring RUG from each secondary cluster was retained in the dereplicated set. For our dataset, 4941 dereplicated RUGs were obtained.

Note that we operate a continuous de-replication workflow. Therefore all 913 of the RUGs (both MetaBAT2 and Hi-C) we previously published have been merged with the newer RUGs and de-replicated. Therefore, whilst some of the previously published RUGs exist in the newer dataset published here, many have been replaced by newer RUGs of higher quality.

Supplementary Data 5 is the average depth for each genome in each sample as calculated by script jgi_summarize_bam_contig_depths from the MetaBAT2[52] (v 2.11.1) package.

*Bioinformatics – metagenomic assignment***.** The output of metagenomic binning is simply a set of DNA FASTA files containing putative genomes. These were all assessed for completeness and contamination using CheckM[35] (v1.0.5). The 4941 best bins were analysed using MAGpy[10], a Snakemake[55] pipeline that runs a series of analyses on the bins, including: CheckM (v1.0.5); prodigal[34] (v2.6.3) protein prediction; Pfam_Scan[56] (v1.6); DIMAOND[12] (v0.9.22.123) search against UniProt TrEMBL; PhyloPhlAn[13] (v0.99); and Sourmash (v2.0.0) search against all public bacterial genomes. The MAGpy results were used to produce a putative taxonomic assignment for each bin as follows:

- If the proportion of proteins assigned to a species is >= 0.9 and the average amino acid identity >= 0.95, assign to species based on DIAMOND results; else
- If sourmash score is >= 0.8 assign to species based on Sourmash results; else
- If PhyloPhlAn probability is high and the level is genus or species, then assign taxonomy based on PhyloPhlAn results; else
- If the proportion of proteins assigned to a genus is >= 0.9 and the average amino acid identity >= 0.9, assign to genus based on DIAMOND results; else
- If PhyloPhlAn probability is high or medium and the level is genus, then assign to genus based on PhyloPhlAn results; else
- If PhyloPhlAn probability is high or medium and the level is family, then assign to family based on PhyloPhlAn results; else
- If the proportion of proteins assigned to a family is >= 0.8 and the average amino acid identity >= 0.6, assign to family based on DIAMOND results; else
- If PhyloPhlAn probability is high or medium and the level is order, then assign to order based on PhyloPhlAn results; else
- If the proportion of proteins assigned to a order is >= 0.6 and the average amino acid identity >= 0.6, assign to order based on DIAMOND results; else
- If PhyloPhlAn probability is high or medium and the level is class, then assign to class based on PhyloPhlAn results; else
- If PhyloPhlAn probability is high or medium and the level is phylum, then assign to phylum based on PhyloPhlAn results; else
- Assign taxonomy based on CheckM lineage

Importantly, at this stage, these are only putative taxonomic assignments. Using these labels, a phylogenetic tree consisting of the RUGs and genomes from the Hungate collection, produced from concatenated protein subsequences using PhyloPhlAn[13] (v0.99), was visually inspected using FigTree (v.1.4.3), iTol[57] (v4.3.1) and GraPhlAn[45] (v0.9.7). Annotations were improved where they could be - for example where MAGpy had only assigned a taxonomy at the Genus level, but that genome clustered closely with a Hungate 1000 genome annotated at the species level, the annotation was updated. The tree was also re-rooted manually at the Bacteria/Archaea branch using FigTree.

Bioinformatics – genome quality and comparative genomics. Genome completeness and contamination was assessed using CheckM (v1.0.5) (see above). tRNA genes were annotated using tRNAscan-SE (v2.0.0) and 16S rRNA genes predicted using barrnap (v0.9). Whole-genome alignments were calculated with MUMmer[58] (v 3.23) using promer to find matches between genomes. Average nucleotide identity was calculated using FastANI (v1.1). The RUGs were

compared to the Hungate collection and our previous dataset using DIAMOND blastp (v0.9.22.123) and MASH[59] (v 2.0) with parameters -k 21 -s 100000.

The rumen supserset was de-replicated using dRep as above, with –sa 0.99 for de-replication at 99% ANI and –sa 0.95 for de-replication at 95% ANI. Overlaps between sets were plotted with UpSetR[60] (v1.3.3). Read classification rats were calculated using Kraken[61] (v0.10.5) with parameters --fastq-input --gzip-compressed --preload --paired.

Bioinformatics – analysis of sheep methane data. Reads from the low and high methane samples from Shi *et al* were assigned to different taxonomic levels of the rumen superset database using Kraken, as described above. The resulting read counts data was used as input into DESeq2 (v1.22.2) for differential analysis. Principal components analysis plots were created using the plotPCA() function within DESeq2, and heatmaps were created using the heatmap.2() function within the gplots package (v3.0.1.1). For strain-level analysis, reads from the low and high methane samples from Shi *et al* were aligned directly to the rumen superset database using BWA MEM (v0.7.15) and the number of primary alignments to each genome was used as input to DESeq2. P-values for all comparisons were calculated by DESeq2 and adjusted for multiple testing[62].

Bioinformatics – rumen census analysis. The average and total depth for each genome in each dataset (Supplementary Data 5) was used as a proxy for abundance in the dataset(s). Kraken (as described above) was used with the rumen superset database to calculate the read abundance of *Proteobacteria* in all samples.

Bioinformatics – assembly and analysis of Nanopore sequence data. The Nanopore reads were extracted and QC-ed using poRe[63,64] (v 0.24), and assembled using Canu[32] (v1.8) with default settings and genomeSize=150Mb. The resulting assembly was analysed using MAGpy[10]. The raw assembly was corrected using both Nanopolish[65] (v 0.10.2) and Racon[66] (v 1.3.1) using Illumina data aligned to the Nanopore assembly with Minimap2 (v 2.12) using short read settings (-x sr). Query vs subject length data were extracted and plotted using ideel (https://github.com/mw55309/ideel). Whole-genome alignments were calculated using MUMmer79 (v 3.23) using promer to find matches between genomes. The three complete nanopore bacterial genomes and their Illumina counterparts were annotated using Prokka[67] (v 1.13.3). The Nanopore assembly was created with a minimum contig length of 1kb, therefore the Illumina assemblies were similarly limited prior to comparison.

*Bioinformatics – proteome analysis***.** Proteins were predicted using Prodigal (v2.6.3) with option –p meta. Using DIAMOND, each protein was searched against KEGG (downloaded on Sept 15th 2018), UniRef100, UniRef90 and UniRef50 (downloaded Oct 3rd 2018), and CAZy (dbCAN2 version, 31/07/2018). The protein predictions were clustered by CD-HIT[68] (v4.7) at 100%, 90% and 50% identity, mirroring similar methods at UniRef.

All protein predictions were searched against the CAZy database using dbCAN2[40] and HMMER[69] (v3.1b2), and polysaccharide utilisation loci (PUL) were predicted for *Bacteroidetes* RUGs using PULpy[43].

References

46. Duthie, C.-A. *et al.* Impact of adding nitrate or increasing the lipid content of two contrasting diets on blood methaemoglobin and performance of two breeds of finishing beef steers. *animal* **10**, 786–

795 (2016).

47. Duthie, C.-A. *et al.* The impact of divergent breed types and diets on methane emissions, rumen characteristics and performance of finishing beef cattle. *animal* **11**, 1762–1771 (2017).

48. Yu, Z. & Morrison, M. Improved extraction of PCR-quality community DNA from digesta and fecal samples. *Biotechniques* **36**, 808–812 (2004).

49. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).

50. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 3 (2013).

51. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).

52. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

53. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

54. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

55. Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

56. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).

57. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).

58. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).

59. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

60. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

61. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

62. Benjamini, Y. & Hochberg, Y. *Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing*. *J. Royal Statist. Soc., Series B* **57**, (1995).

63. Watson, M. *et al.* poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* **31**, 114–5 (2015).

64. Stewart, R. D. & Watson, M. poRe GUIs for parallel and real-time processing of MinION sequence data. *Bioinformatics* **33**, 2207–2208 (2017).

65. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).

66. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

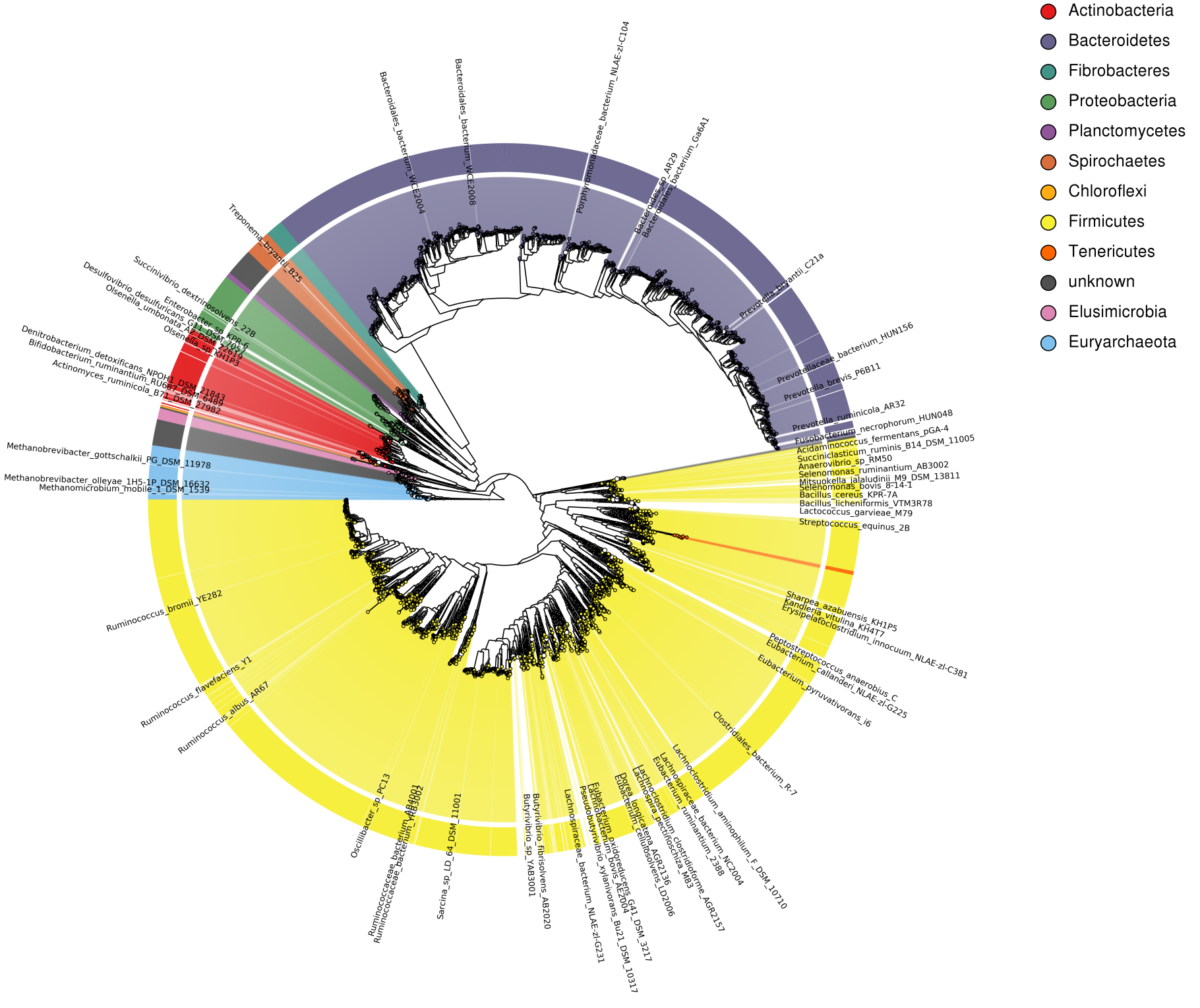
67. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

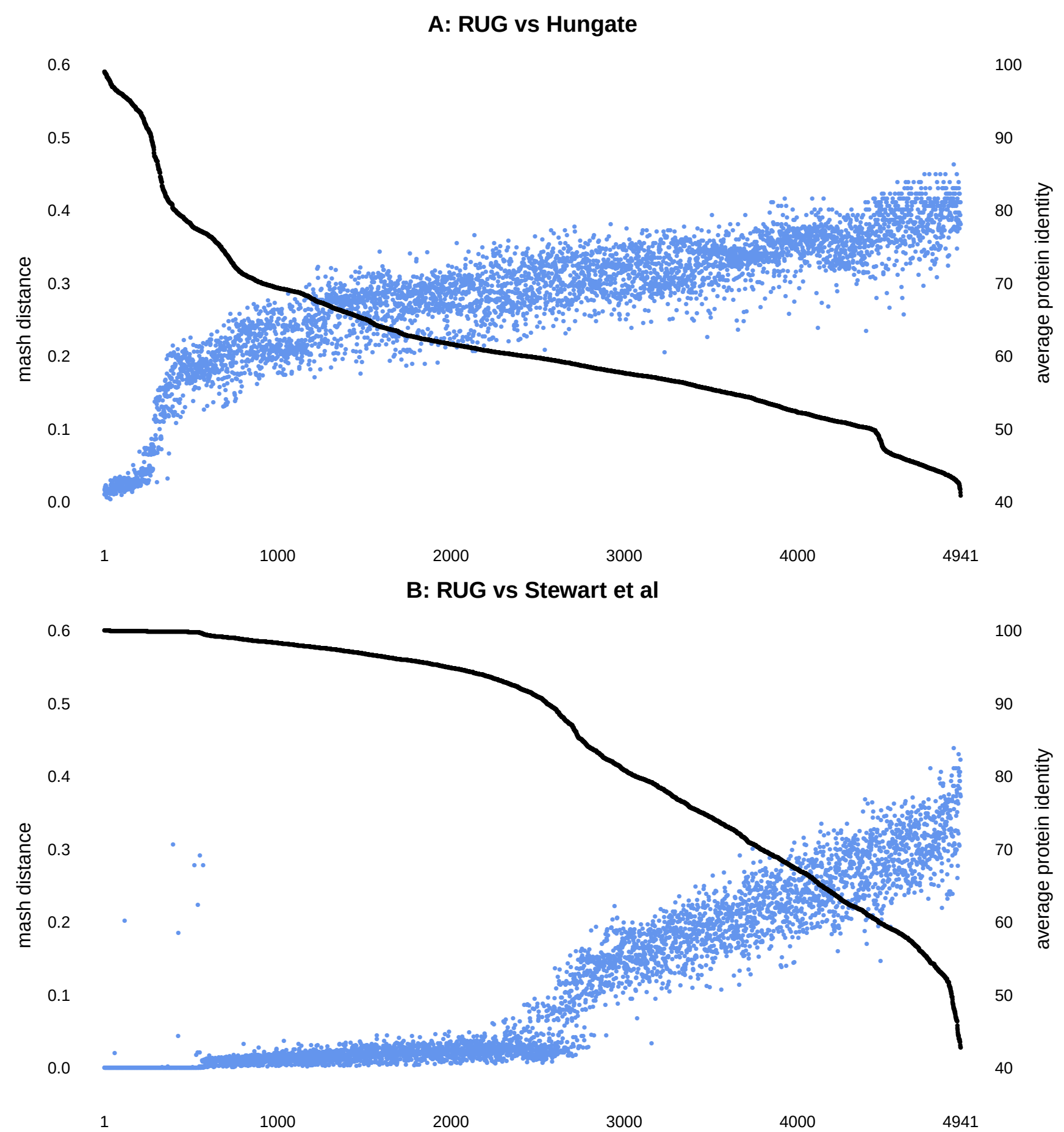
68. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

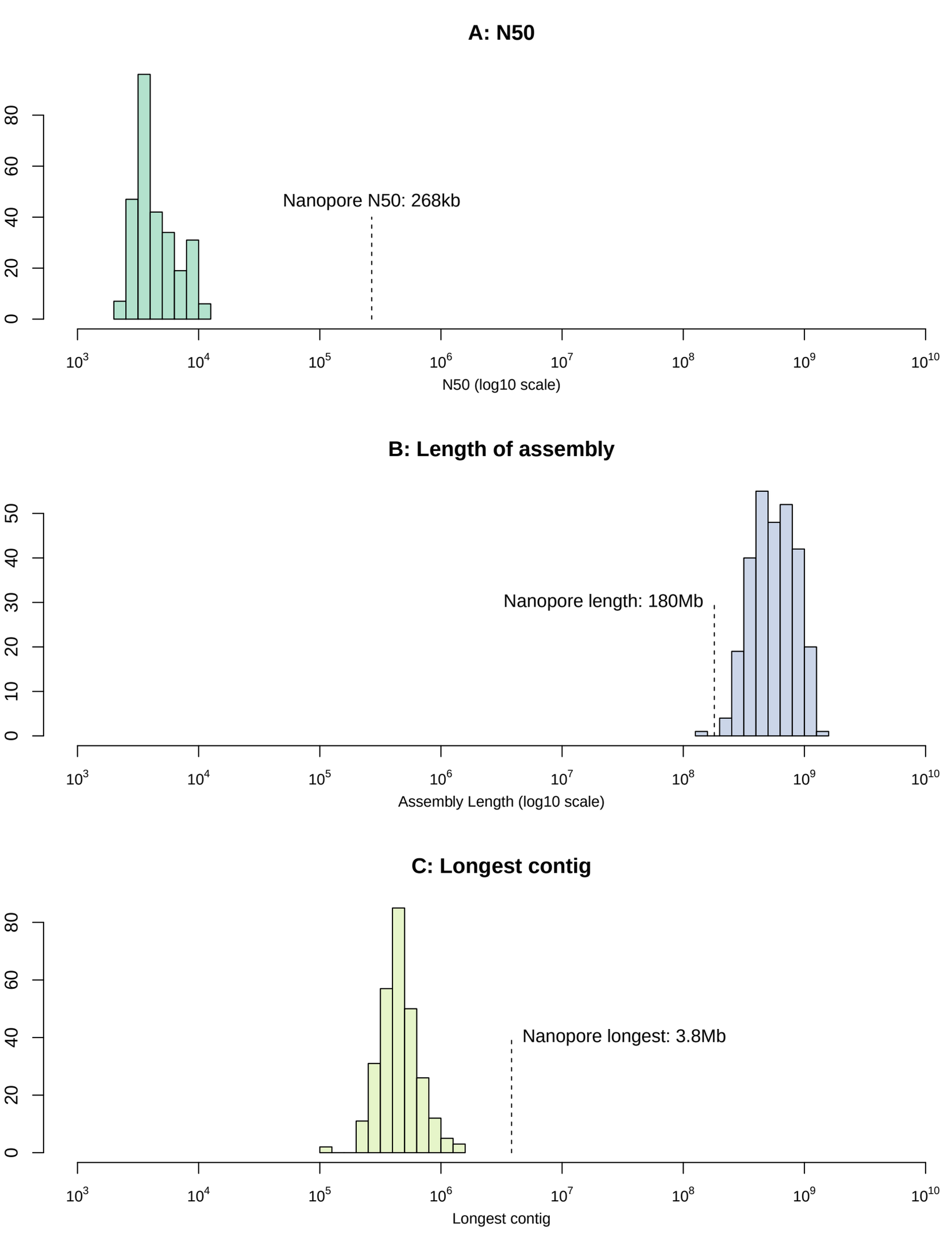
69. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121–e121 (2013).

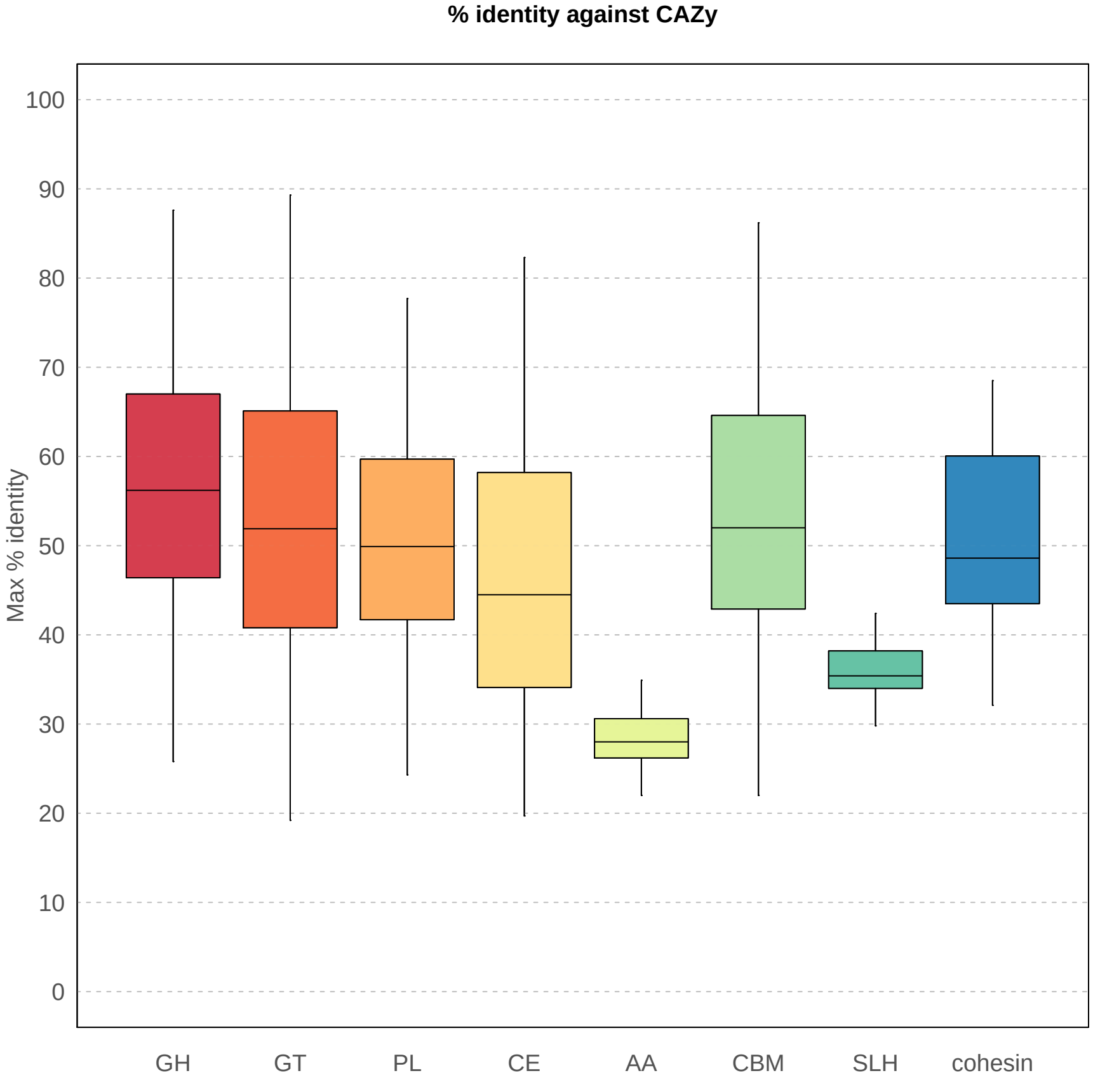
Author contributions

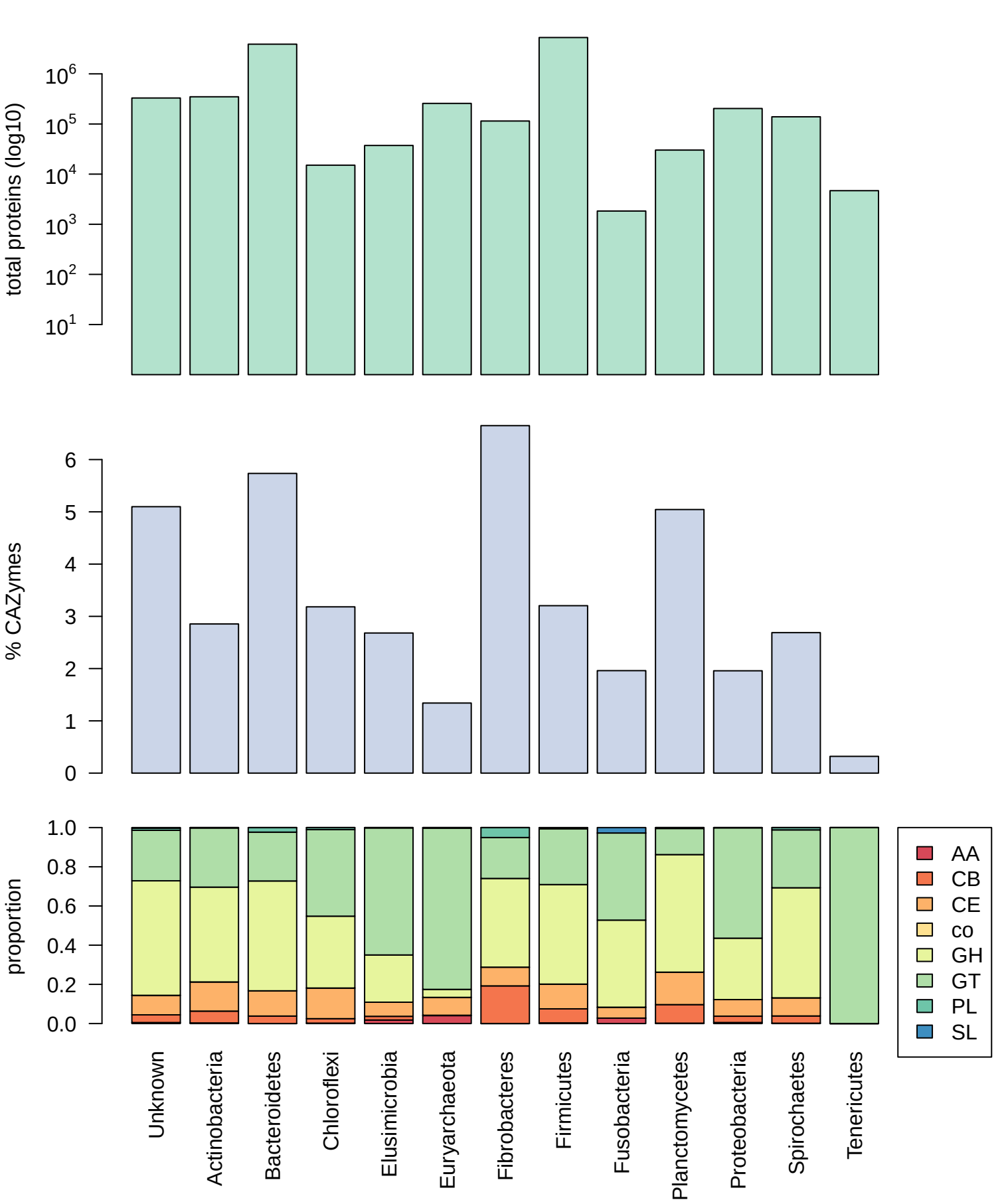
MW, RR and AWW conceived of the study and supervised the project. RDS and MW carried out all bioinformatics work on the Illumina data and AW carried out all bioinformatics work on the Nanopore data. MA carried out all laboratory work, except for DNA clean up, Nanopore library prep and Nanopore sequencing which was done by AW. All of the authors contributed ideas, co-wrote the paper, and reviewed and approved the manuscript.


Actinobacteria
Bacteroidetes
Fibrobacteres
Proteobacteria
Planctomycetes
Spirochaetes
Chloroflexi
Firmicutes
Tenericutes
unknown
Elusimicrobia
Euryarchaeota
Bacteroidales_bacterium_WCE2004
Bacteroidales_bacterium_WCE2008
Porphyromonadaceae_bacterium_NLAE-zl-C104
Bacteroides_sp_AR29
Bacteroidales_bacterium_Ga6A1
Prevotella_bryantii_C21a
Prevotellaceae_bacterium_HUN156
Prevotella_brevis_P6B11
Prevotella_ruminicola_AR32
Fusobacterium_necrophorum_HUN048
Fusobacterium_fermentans_pGA-4
Acidaminococcus_ruminis_B14_DSM_11005
Succiniclasticum_ruminis_B14_DSM_11005
Anaerovibrio_sp_RM50
Selenomonas_ruminantium_AB3002
Mitsuokella_jalaludinii_M9_DSM_13811
Selenomonas_bovis_8-14-1
Bacillus_cereus_KPR-7A
Bacillus_licheniformis_VTM3R78
Lactococcus_garvieae_M79
Streptococcus_equinus_2B
Sharpea_azabuensis_KH1P5
Kandleria_vitulina_KH4T7
Erysipelatoclostridium_innocuum_NLAE-zl-C381
Pentostreptococcus_anaerobius_C
Eubacterium_callanderi_NLAE-zl-G225
Eubacterium_pyruvativorans_I6
Clostridiales_bacterium_R-7
Lachnoclostridium_aminophilum_F_DSM_10710
Lachnospiraceae_bacterium_NC2004
Eubacterium_ruminantium_2388
Lachnoclostridium_clostridioforme_AGR2157
Lachnospira_pectinoschiza_M83
Eubacterium_cellulosolvens_LD2006
Lachnoclostridium_AGR2136
Dorea_longicatena_AGR2135
Eubacterium_oxidoreducens_G41_DSM_3217
Pseudobutyrivibrio_xylanivorans_Bu21_DSM_10317
Lachnospiraceae_bacterium_NLAE-zl-G231
Butyrivibrio_sp_YAB3001
Butyrivibrio_fibrisolvens_AB2020
Sarcina_sp_LD_64_DSM_11001
Ruminococcaceae_bacterium_YAB3002
Ruminococcaceae_bacterium_AB4001
Oscillibacter_sp_PC13
Ruminococcus_albus_AR67
Ruminococcus_flavefaciens_Y1
Ruminococcus_bromii_YE282
Methanobrevibacter_gottschalkii_PG_DSM_11978
Methanobrevibacter_olleyae_1H5-1P_DSM_16632
Methanomicrobium_mobile_1_DSM_1539
Actinomyces_ruminicola_B71_DSM_27982
Bifidobacterium_ruminantium_RU687_DSM_6489
Denitrobacterium_detoxificans_NPOH1_DSM_21843
Olsenella_sp_KH1P3
Olsenella_umbonata_A2_DSM_22619
Desulfovibrio_desulfuricans_G41_DSM_14057
Enterobacter_sp_KPR-6
Succinivibrio_dextrinosolvens_22B
Treponema_bryantii_B25


A: RUG vs Hungate
mash distance
average protein identity
0.6
0.5
0.4
0.3
0.2
0.1
0.0
100
90
80
70
60
50
40
1
1000
2000
3000
4000
4941
B: RUG vs Stewart et al
mash distance
average protein identity
0.6
0.5
0.4
0.3
0.2
0.1
0.0
100
90
80
70
60
50
40
1
1000
2000
3000
4000
4941


A: N50
Nanopore N50: 268kb
0
20
40
60
80
10^3
10^4
10^5
10^6
10^7
10^8
10^9
10^{10}
N50 (log10 scale)
B: Length of assembly
Nanopore length: 180Mb
0
10
20
30
40
50
10^3
10^4
10^5
10^6
10^7
10^8
10^9
10^{10}
Assembly Length (log10 scale)
C: Longest contig
Nanopore longest: 3.8Mb
0
20
40
60
80
10^3
10^4
10^5
10^6
10^7
10^8
10^9
10^{10}
Longest contig


% identity against CAZy
Max % identity
100
90
80
70
60
50
40
30
20
10
0
GH
GT
PL
CE
AA
CBM
SLH
cohesin


total proteins (log10)
10^6
10^5
10^4
10^3
10^2
10^1
% CAZymes
0
1
2
3
4
5
6
proportion
0.0
0.2
0.4
0.6
0.8
1.0
Unknown
Actinobacteria
Bacteroidetes
Chloroflexi
Elusimicrobia
Euryarchaeota
Fibrobacteres
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
Spirochaetes
Tenericutes
AA
CB
CE
co
GH
GT
PL
SL